TITLE: Multilingual Text-Independent Speaker Identification

MULTILINGUAL TEXT-INDEPENDENT SPEAKER IDENTIFICATION

Geoffrey Durou

Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez
B-7000 Mons, Belgium
Email: durou@tcts.fpms.ac.be

ABSTRACT

In this paper, we investigate two facets of speaker recognition : cross-language speaker identification and same-language non-native text-independent spea-ker identification. In this context, experiments have been conducted, using standard multi-gaussian mo-deling, on the brand new multi-language TNO corpus. Our results indicate how speaker identification performance might be affected when speakers do not use the same language during the training and testing, or when the population is composed of non-native speakers.

1. INTRODUCTION AND MOTIVATION

Speaker recognition systems working in text independent (TI) mode have been characterized by their flexibility but also by their insecure aspect. Indeed, the non-imposing of words or sentences can lead to the breaking of the system if the voice of an authorized person is pre-recorded.

However, text-independent speaker identification systems are involved in many applications. That is the reason why many efforts have been developed in order to improve text-independent speaker recognition methods. For the last decade, the technology in this field has achieved significant progress. Now, these techniques can be used in real conditions, for that the application field be well defined.

Nowadays, more and more users of such systems are polyglot. So, if we do not have a priori know-ledge of the mother tongue of the talker - or at least the tongue he used during the training - and if we can not apply any language identification system, then it is possible to perform speaker identification in a language different from the one used during training. Let us note that no restriction about the tongue would still increase the flexibility of the system. However, the system may still impose one specific tongue. Since, it should be open to all users, we can easily imagine that any given language might differ from the native language of some of the users.

In order to start a descriptive study on (a) the cross-language and (b) the same non-native language effects on speaker recognition performance, we carried out some text-independent speaker identification experiments on a subset of 57 speakers extracted from the TNO multi-language database. Our system is based on the standard GMM technique, which has already been successfully used by the past for TI speaker recognition [3] [2] [4].

In section 2 we present in detail the TNO corpus and our identification system. The speaker identification experiments are described in section 3, which is subdivided into three items : (a) native speaker identification, acting as reference experiment; (b) cross-language speaker identification; (c) non-native same-language speaker identification. Results are then discussed and, in particular, cross-language spea-ker identification results are compared to performance recently obtained on the POLYCOST telephone speech corpus [5] [1].

2. EXPERIMENTAL SETUP

2.1. Database

Speech material for our experiments was taken from the new Dutch TNO corpus. This database consists in 82 Dutch speakers. All of them were prompted to pronounce 10 sentences in four different languages : Dutch, English, French, and German. All the sentences were read from a computer screen in a anechoic silent recording room. Given one language, the first five sentences are common for all speakers, while the others differ from one speaker to another.

We decided to accomplish the identification tests over all the speakers for whom speech data in the four tongues are available. So we conducted our experiments on a subset of 57 speakers (68 % males and 32 % females).

The first 5 utterances (per language identical for all speakers) were used for the training, while the other 5 sentences (per language and per speaker unique) were reserved to the identification tests.

In our experiments, we have systematically considered four different training durations (10 s, 15 s, 20 s, and 25 s) and five different testing durations (5 s, 10 s, 15 s, 20 s, and 25 s).

2.2. Feature Extraction

Speech recordings were sampled at 16 kHz. Analysis windows consisted of 512 samples taken every 16 ms. After pre-emphasis (factor 0.95) and application of a Hamming window, 10 autocorrelation LPC coefficient were computed and transformed into 12 cepstral coefficients. Finally, training and testing features consist only of 12 cepstral coefficients : neither the energy, nor dynamic information (delta coefficients), nor the pitch were used. No cepstral mean subtraction was applied.

2.3. Speaker Model

Our speaker identification system is based on the statistical modeling by Gaussian mixtures [3] [2] [4]. Each mixture is composed of 12 Gaussian distributions, with diagonal covariances matrices.

3. EXPERIMENTS

3.1. Native speaker Identification

First of all, let us carry out a preliminary experiment, considering both training and test phases in the mother tongue of the speakers. This might be seen, in the context of this paper, as the reference experiment.

Let us remind once again that for these experiments and all the experiments that will follow, we shall systematically choose the five sentences per language identical for the training, and the other five per language and per speaker unique for the identification tests.

The identification error rates for various training and testing durations are given hereafter in Figure 1.

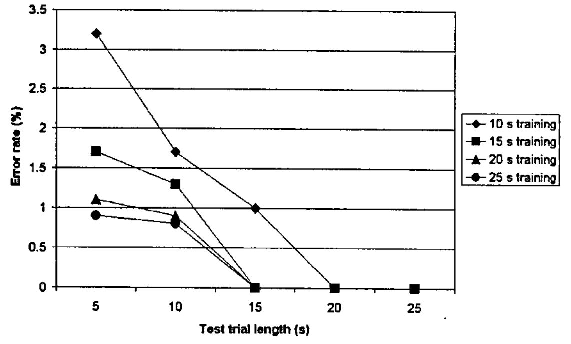

Figure 1: *Identification error rates over 57 native speakers of Dutch as a function of test trial length for various training conditions*

We can notice at this point that the closed set speaker identification rate reaches 100 % for a 20 second testing duration and more, whatever the training duration considered.

3.2. Cross-language speaker identification

It would now be interesting to measure the impact of language on our speaker recognition system.

For that purpose, we conduct an experiment characterized by the use of different languages during the training and the test : models are trained on native speech (i.e. Dutch), while identification tests are made successsively on non-native speech (successively English, French, and German).

Results for different training and testing durations are reported in Figure 2, Figure 3 and Figure 4 below.

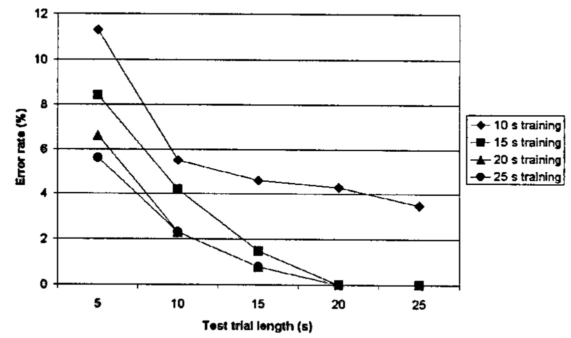

Figure 2: *Cross-language speaker identification error rates (Dutch / English) over 57 Dutch speakers as a function of test trial length for various training conditions.*

For values of training and testing durations large enough, we are still able, in the case Dutch/English, to reach the maximal performance.

On the contrary, we are unable to reach a 100 % identification rate in the case Dutch/French, given our proposed training and testing conditions.

When German is used for the test, error rates seem to converge to about 2 %.

Similar experiments have been recently conducted on a telephone speech database [1]. In this context, cross-language speaker identification tests on a set of 111 speakers showed that the performance degradation induced by

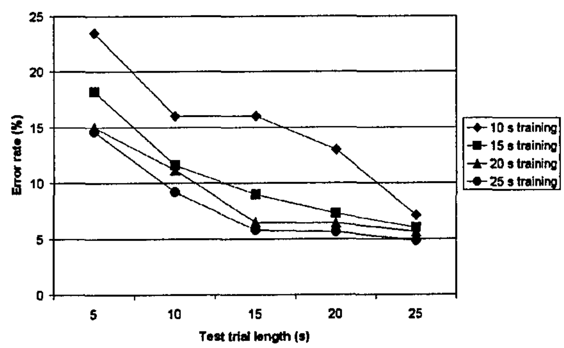

Figure 3: *Cross-language speaker identification error rates (Dutch / French) over 57 Dutch speakers as a function of test trial length for various training conditions.*

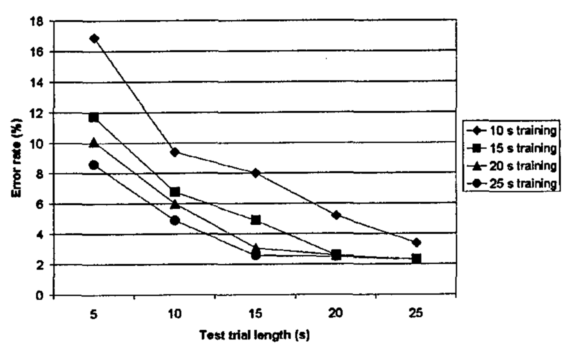

Figure 4: *Cross-language speaker identification error rates (Dutch / German) over 57 Dutch speakers as a function of test trial length for various training conditions.*

the use of a non-native tongue for the test did not exceed 1 % (relatively to the use of the native tongue for the test) in the case of a speaker identification system based on a vector quantization technique. We justified this very restricted difference by the fact that spectral characteristics of the speaker speech is not importantly modified as he speaks a second language. This corroborated another study which has shown that people who learn a second language at an advanced age (> 10 years old), instead of learning new phonemes, substitute phonemes from their native language and impose the rythm of this native language as they speak a non-native language [8]. Let us also mention that this conclusion was consolidated by an experiment described in [6] and which showed that the spectrum difference, measured by Kullback's divergence, on English and Japanese words pronounced by bilingual speakers was very small.

Here, in the case of maximal training and testing durations, we observe that the degradation easily exceeds 1 % in the cases Dutch-French (4.8 %) and Dutch-German (2.3 %) even though the population size is more restricted. However, we must be aware that, first, the maximal training duration is here of 25 seconds, whereas each training session lasted about 90 seconds in the previous work. Secondly, our identification system is now based on statistical modeling by Gaussian mixtures. These two points make it difficult to compare in the absolute results from these experiments.

3.3. Non-native speaker identification

Let us finally consider a last set of experiments conducted on non-native talkers. We conducted three sets of experiments characterized by the use of same non-native language during the training and the test : models were trained and identification tests were made on non-native speech (successively English, French, and German).

Once again, we report separately results on English, French, and German speech in Figure 5, Figure 6, and Figure 7, for different training and testing durations.

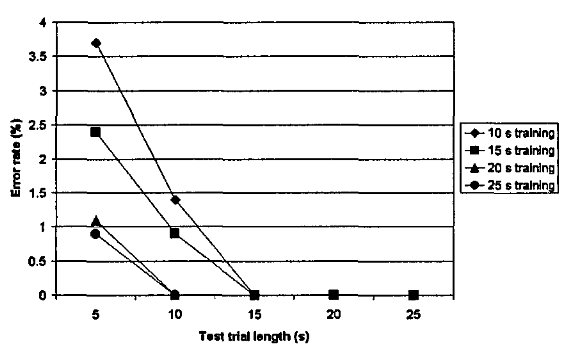

Figure 5: *Identification error rate over 57 non-native speakers of English as a function of test trial length for various training conditions.*

When English is chosen as non-native language, we see that there is no big difference between these plots and the reference plots. Surprisingly enough, the system performs sometimes better when this non-native language is employed.

We may reiterate the same observation if German is used. However, our system performs slightly worse if French is employed.

Globally, as expected, we observe through these experiments that even if non-native speakers use the phonetic and prosodic patterns of their first language, the identification scores are not really affected.

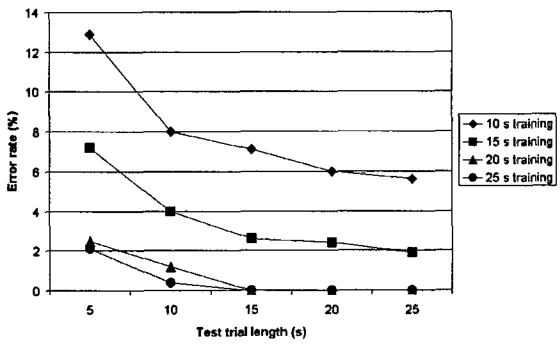

Figure 6: *Identification error rate over 57 non-native speakers of French as a function of test trial length for various training conditions.*

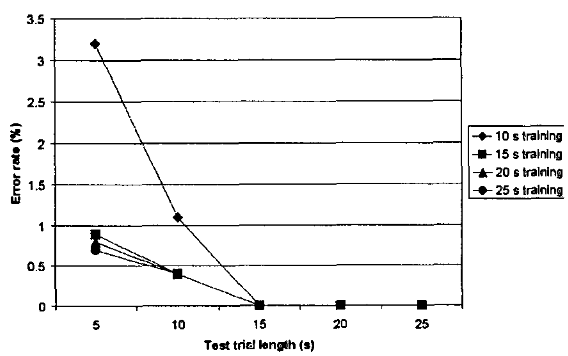

Figure 7: *Identification error rate over 57 non-native speakers of German as a function of test trial length for various training conditions.*

Major aspects that can make non-native speech deviate from native speech are notably fluency, word stress, and intonation [7]. Although these factors might be responsible of a score degradation in the cross-language case, we can easily understand that they have a much more restricted effect on these last experiments. In particular, if a non-native talker tends to speak more slowly during the training, he will also tend to speak roughly the same way for the tests, because the language is the same. This point should explain partly why the identification scores are not so affected.

4. CONCLUSION

The purpose of this paper was to describe and carry out multi-lingual speaker identification experiments on the TNO database made of native speakers of Dutch, and to comment on the results. Various training and testing durations were considered.

We first carried out a preliminary set of experiments (what we considered as being the baseline experiments) where both training of the speakers models and the identification tests were made on their mother tongue (i.e. Dutch). Then, regarding to our baseline results, we have measured the evolution of our speaker identification system performance when (a) different languages are used during the training and the tests; (b) a same non-native language is used both for the speakers models training and the identification tests. Three non-native languages were tested : English, French, and German.

We also pointed out and partly justified the discordance between the conclusions about the effect on the language if the performance degradation is measured on the microphone TNO corpus or on the telephone POLYCOST database.

5. REFERENCES

[1] G. Durou, F. Jauquet, "Cross-Language Text-Independent Speaker Identification", Proc. European Conference on Signal Processing (EUSIPCO'98), vol 3, pp 1481-1484, September 1998, Rhodes, Greece.

[2] D. A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification", PhD Thesis, Georgia Institute of Technology, 1992.

[3] G. McLachlan and K. Basford, "Mixture Models : Inference and Applications to Clustering", Marcel Dekker, 1998.

[4] D. Titterington, A. Smith, and U. Markov, "Statistical Analysis of Finite Mixture Distributions", John Wiley and sons, 1985.

[5] The European COST 250 action entitled "Speaker Recognition in Telephony", Information can be found on the web page : http://circhp.epfl.ch/polycost/

[6] M. Abe and K. Shikano, "Statistical analysis of bilingual speakers's speech for cross-language voice conversation", J. Acoust. Soc. Amer., Vol 90, pp 76-82, July 1991.

[7] C. Cucchiarini, H. Strik, and L. Boves, "Automatic evaluation of Dutch pronunciation by using speech recognition technology", Proc IEEE ASRU, Santa Barbara, Dec 1997.

[8] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech", Proc ICSLP'96, Philadelphia, pp 1457-1460, 1996.